Prompts to Proxies

Emulating Human Preferences via a Compact LLM Ensemble

Bingchen Wang[1] · **Zi-Yu Khoo**[2] · **Jingtan Wang**[3]

[1]*Independent Researcher* · [2]*AI Singapore* · [3]*School of Computing, National University of Singapore*

NEURAL INFORMATION PROCESSING SYSTEMS

SYDNEY · NEURIPS 2026 · MAIN CONFERENCE

PROJECT PAGE AND CODE: SCAN THE CODES, TOP RIGHT

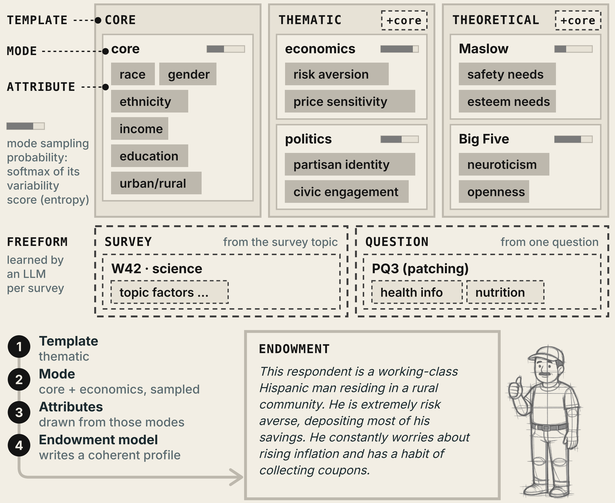
ARXIV 2509.11311

Project page *Code*

Q1 MECHANISM

What exactly is a proxy?

An LLM prompted with an **endowment**: a profile written from attributes drawn out of the **attribute bank**.



1. **Template** thematic
2. **Mode** core + economics, sampled
3. **Attributes** drawn from those modes
4. **Endowment model** writes a coherent profile

ENDOWMENT

This respondent is a working-class Hispanic man residing in a rural community. He is extremely risk averse, depositing most of his savings. He constantly worries about rising inflation and has a habit of collecting coupons.

Attributes are **control handles**, not profiles matched to real people.

Q2 THEORY

Why an ensemble?

A survey aggregates latent preferences p_i; a **functional basis** of proxies $\hat{p}_j$ can reconstruct it with no one-to-one match:

$$r_{\text{pop}} = \sum_i w_i \, h(p_i) \qquad r_{\text{pop}} = \sum_j \tilde{w}_j \, f(\hat{p}_j)$$

Representative agent	**Demographic conditioning**	**Functional basis** OURS
one model preference	profile → proxy	weighted ensemble

What a representative agent can learn, an ensemble can; some populations may be learnable **only by an ensemble**; if demographic conditioning works, the ensemble is **not unique**. Learnable iff $\text{conv}\, h(\mathcal{P}) \subset \text{conv}\, f(\widehat{\mathcal{P}})$.

Q3 SIMULATION & ABLATION

Does the selection work?

- In simulations, lasso **recovers the true agents**, minorities included.
- With them removed, it still **reconstructs** from few proxies.
- **Low-entropy pools fail** ⇒ Stage 1 targets response diversity.

ATP W42 · TEST MSE, LOWER IS BETTER

Simple averaging	.0332
P2P regression	.0087

Same proxy pool; only the aggregation differs.

IN NUMBERS

41% mean per-wave gain in test MSE over the best baseline, 14 ATP waves

0.014 average test MSE, against .025 (Vanilla) and .029 (PERSONA)

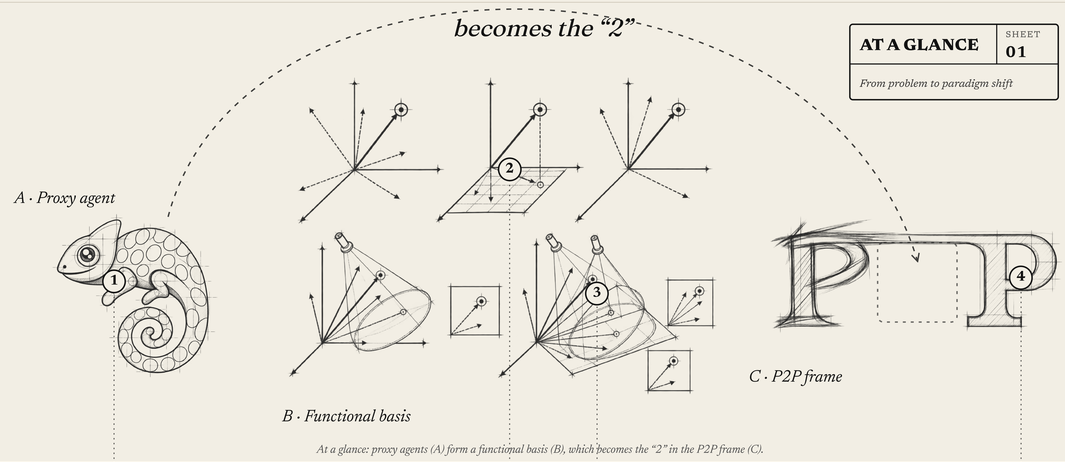
≈$0.8 average cost per ATP wave

<3% of the SFT model's training data, on the WVS topic shift

THE IDEA

Don't profile individuals. Reconstruct the population.

A basis of LLM proxies, weighted by L1-regularised regression, recovers a population's survey responses with no fine-tuning and no individual-level data: a 41% mean per-wave gain over the best baseline across 14 ATP waves, at about $0.8 per survey.



AT A GLANCE SHEET 01

From problem to paradigm shift

At a glance: proxy agents (A) form a functional basis (B), which becomes the "2" in the P2P frame (C).

01 **PROBLEM**

LLM proxies stand in for human subjects, but conclusions hold (**external validity**) only if they reflect the real population's preferences.

02 **WHY RETHINK IT**

Fine-tuned models tend to collapse to central tendencies; profiling needs **sensitive data** and doesn't guarantee **distributional fidelity**.

03 **IDEA**

Preference reconstruction: build a response-expressive **basis** of proxies, then recover the population by **weighted aggregation**.

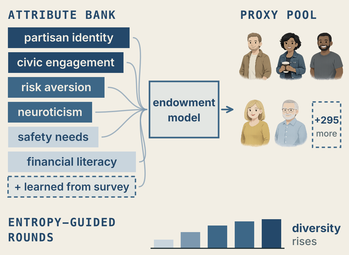
04 **RESULT**

No fine-tuning, no individual-level data: test MSE **0.014** on 14 American Trends Panel (ATP) waves, a **41%** mean per-wave gain over the best baseline, **≈$0.8 per survey**.

HOW P2P WORKS

STAGE 1 · GENERATE

Build a diverse proxy pool

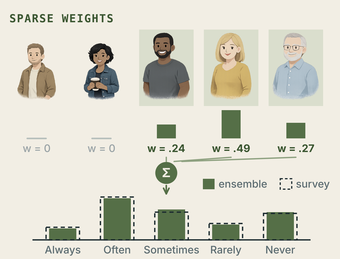


An **endowment model** turns sampled attributes into coherent profiles; **entropy-guided** sampling favours response diversity. No human answers are used.

$$H_i(\mathcal{A}) = \frac{-\sum_k p_{ik} \log_2 p_{ik}}{\log_2 K_i}$$

STAGE 2 · AGGREGATE

Select a compact ensemble

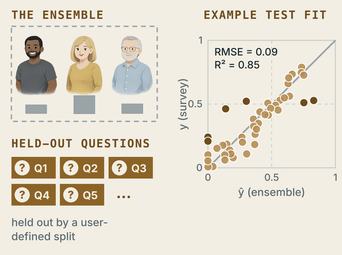


Constrained lasso or elastic net selects and weights proxies so their aggregate matches the survey on training and validation questions.

$$\min_{\mathbf{w}} \sum_{i,k} \left(r_{ik} - \sum_j w_j d_{j,ik}\right)^2 + \alpha\, \Omega(\mathbf{w})$$

THEN · VALIDATE

Test on unseen questions



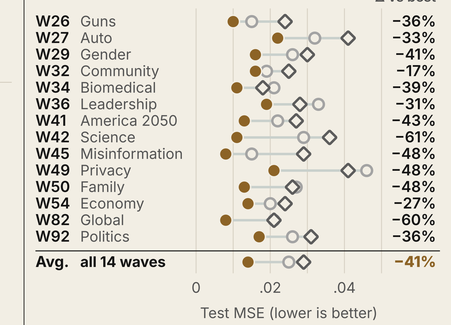
A **user-defined split** (70 / 15 / 15 in our experiments) holds out test questions that neither stage sees.

test MSE on unseen questions

Q4 RESULTS

How well does it work?

○ Vanilla ◇ PERSONA ● P2P (ours)



Shared Stage 2, so the gap reflects **proxy-pool quality** (entropy .53 vs .43–.45).

Δ: per-wave gain over the best baseline; 3 runs.

Q5 COMPARISON

Versus fine-tuning (SFT)?

WVS topic shift	Train rows	1–JSD ↑	MCPA ↑
Cao et al. (2025): SFT on 46 countries	≈8,400	0.78	0.43
P2P, per locale	**172–185**	**0.71–0.75**	**0.49–0.64**

With under **3%** of the training data, P2P **matches or beats** SFT on MCPA and comes close on 1–JSD.

MCPA: majority-class prediction accuracy; JSD: Jensen–Shannon divergence; WVS: World Values Survey.

Q6 APPLICATIONS

Use cases

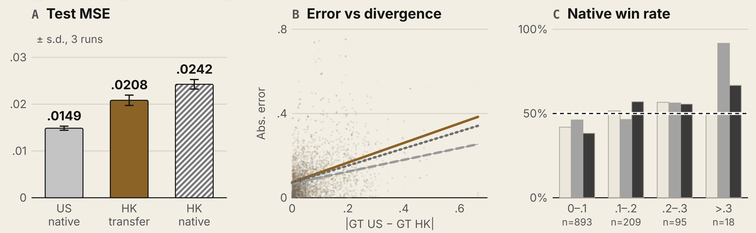
- **Zero-demographic benchmark:** how much of a survey is recoverable with no demographic or private data; richer methods should beat it.
- **Imputing unasked questions:** calibrate on questions shared across waves; on WVS W6/W7 (72 anchors), test MSE falls 12.1% vs. persistence.

Scope: generalises to unseen questions within a survey, not across surveys.

Q7 TRANSFER

Must proxies match the locale?

HK-transfer refits only Stage 2 on *US-generated* proxies; US–HK r = .848.



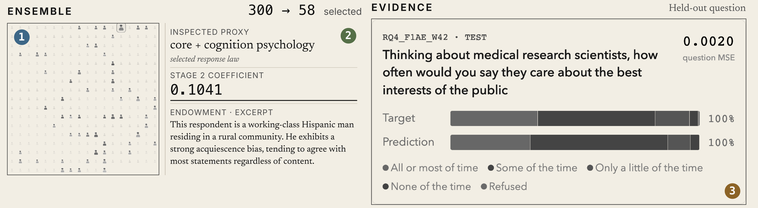
- **Expressiveness wins on average:** HK-transfer .0208 vs HK-native .0242.
- Error grows with US–HK divergence: an observable diagnostic of **learnability limits**.
- **Localisation helps in the tails:** native wins more often on divergent questions.

Q8 IN PRACTICE

What does a run look like?

P2P Studio — Local research workspace

Inspect the ensemble. Check its predictions.



1 **Generate.** 300 proxies, one figure each; dimmed ones are not kept.

2 **Aggregate.** Lasso keeps 58; the inspected one has the largest positive weight, 0.104.

3 **Validate.** On a held-out question, prediction tracks survey (MSE .002).

One run, end to end: a broad pool, a sparse ensemble, a check on an unseen question.

Paper run: ATP W42 · lasso · Gemini-2.0-flash · `pip install -e . && p2p ui`

References. Castricato et al. (2025), PERSONA · Santurkar et al. (2023), OpinionsQA · Cao et al. (2025), distribution-calibrated SFT · Pew Research Center, American Trends Panel · World Values Survey, Wave 7